(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－167001
（P2001－167001A）
(43)公開日　平成13年６月22日(2001.6.22)

(51)Int.Cl.[7]	識別記号	ＦＩ	テーマコード（参考）
Ｇ０６Ｆ　12/16	３１０	Ｇ０６Ｆ　12/16	３１０Ｑ
			３１０Ｊ
	３２０		３２０Ｆ
			３２０Ｌ

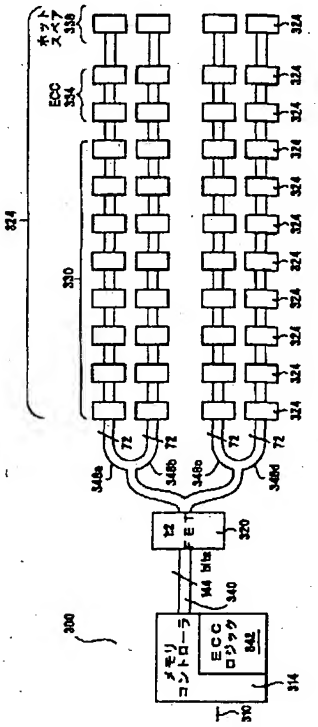
審査請求　未請求　請求項の数１　ＯＬ　（全 11 頁）

(21)出願番号　特願2000－312778(P2000－312778)
(22)出願日　平成12年10月13日(2000.10.13)
(31)優先権主張番号　０９／４３０３６３
(32)優先日　平成11年10月28日(1999.10.28)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　398038580
ヒューレット・パッカード・カンパニー
ＨＥＷＬＥＴＴ－ＰＡＣＫＡＲＤ　ＣＯＭＰＡＮＹ
アメリカ合衆国カリフォルニア州パロアルト　ハノーバー・ストリート　3000
(72)発明者　マイケル・ビー・レイハム
アメリカ合衆国95033カリフォルニア州ロス・ガトス、クヌース・ロード　18219
(72)発明者　ジェームス・ジー・マティオス
アメリカ合衆国95070カリフォルニア州サラトガ、ウッドサイド・ドライブ　12541
(74)代理人　100081721
弁理士　岡田　次生

(54)【発明の名称】　自己回復するメモリ構成

(57)【要約】
【課題】　複雑なデータ修復方法を必要とせず、複数チャネル上の複数エラーに対するエラー検出および訂正を行うと共に、障害の発生したメモリのホットスワッピングもサポートする、メモリシステム構成を提供する。
【解決手段】　本メモリ構成には、メモリコントローラに電気的に結合されているＣＰＵバスと、スイッチング手段に接続されたメモリコントローラと、複数のメモリモジュールに電気的に接続されたスイッチとを含む。複数のメモリモジュールは、少なくとも一つのデータ用のメモリモジュールと、少なくとも一つのデータ訂正用のメモリモジュールと、少なくとも一つのホットスペアモジュールとを含む。データ用メモリモジュールと、エラー訂正用メモリモジュールと、ホットスペアメモリモジュールとは並列に接続されている。



【特許請求の範囲】
【請求項１】　メモリコントローラに電気的に結合されたＣＰＵバスと、
前記メモリコントローラに電気的に結合されたデータバスと、
少なくとも１つのデータ記憶用メモリモジュールと、
少なくとも１つのスペアメモリモジュールと、を具備し、
前記少なくとも１つのデータ記憶用メモリモジュールおよび前記少なくとも一つのスペアメモリモジュールは、前記データバスに並列に電気的に結合されており、前記少なくとも１つのデータ記憶用メモリモジュールで発生する訂正可能なエラーに応答して、前記少なくとも１つのデータ記憶用メモリモジュールからのデータが前記スペアメモリモジュールにマップされるメモリ構成。

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、メモリ障害後に自己回復することができるサーバ用のメモリシステムに関する。
【０００２】
【従来の技術】利用可能なサーバメモリシステムのサイズは、時と共に増大し続けており、現サーバメモリシステムは、６４ギガバイト以上であることも多い。メモリシステムのサイズが増大するに従い、メモリビットの障害、すなわちメモリシステムに障害が発生する可能性も増加している。メモリシステム障害は、一時的エラーまたは永続的エラーの両方で起こり得る。一時的エラーは、一般に、アルファ粒子によって起こる。永続的エラーには、一般に、メモリセルウォール障害、行または列デコーダ障害、状態機械障害、またはデュアルインラインメモリモジュール（以下、「ＤＩＭＭ」とする）の印刷回路基板とシステム印刷回路基板間の機械的インタフェースの故障といった他の破局故障などがある。
【０００３】メモリシステムの障害を防止するために、種々の形態のメモリ検出および訂正プロセスが開発されている。一般的に使用されているシステムのうちの１つでは、エラーを検出するパリティビットを使用する。データを受信すると、予期される値に対してデータのパリティがチェックされる。データが予期されるパリティ値（奇数または偶数）と一致しない場合は、エラーが発生したと判断される。この方法は単一のビットエラーを判断するには機能するが、複数のビットエラーを判断するには必ずしも適切に作用するものではない。更に、最も簡単なパリティシステムは、データエラーを訂正する仕組みを有していない。
【０００４】Gaskinsらに対する米国特許第5,463,643号は、複数エラーを検出できるパリティシステムを提供する。Gaskinsらの構成は、データエラーを訂正する機構を含むRAMBUSメモリシステムと共に動作するように設計されている。ＲＡＭＢＵＳ仕様には、１６ビットデータパスとパリティビットと種々の制御信号、電源信号および接地信号とを含むメモリチャネル仕様が含まれる。
【０００５】図１は、直列に接続された複数のメモリモジュールを有する一般的なＲＡＭＢＵＳメモリ構成１００の高レベル部分ブロック図を示す。図１には、メモリチャネル１１２およびＣＰＵバス１１４に電気的に接続されたメモリＩ／Ｏコントローラ１１０が示されている。メモリチャネル１１２は、ＲＡＭＢＵＳ構成仕様に従って直列に接続された複数のメモリモジュール１１６ａ，１１６ｂ，…１１６ｎに電気的に接続されている。
【０００６】Gaskinsらによって述べられているＲＡＭＢＵＳ仕様に起こり得る問題の１つとして、パリティチェックがメモリエラーの検出用にのみ供されており、エラー訂正用には限定的な提供しかされていない、ということがある。Gaskinsらの構成はデータエラーの訂正はできるが、そのエラー訂正能力は限定的である。図２を参照すると、４つのデータチャネル２１０ａ～２１０ｄおよび１つのパリティチャネル２１０ｅを有する、Gaskinsらのデータチャネルおよびパリティチャネルの高レベル部分ブロック図が示されている。Gaskinsらの構成は単一チャネル上の単一または複数のエラーを訂正することができるが、出願人は、この構成では複数チャネル上で２つ以上のエラーを検出することができないと考えている。例えば、チャネル２１０ａで完全な（すべてのビット）障害が発生した場合は、データを再構成することができる。しかしながら、異なるチャネル上で２つのビット障害が発生した（例えば、チャネル２１０ａ上で１つのビット障害が発生し、チャネル２１０ｂ上で１つのビット障害が発生する）場合、エラーの検出はできず、データを再構成できない。
【０００７】Gaskinsらの構成における更なる問題は、メモリＩ／Ｏコントローラ上のＩ／Ｏピン数に関連する。メモリ記憶要件の増加につれ、メモリコントローラに接続されたメモリモジュールの数も増加する。直列アーキテクチャ構成では、メモリチャネルの数を１つ増やすと、メモリコントローラに直接接続されたメモリチャネルの数も１つ増える。各メモリチャネルを追加するごとに入力ピンを追加する必要があり、システムピン数が増加するので、これは問題となる。この問題は、特に大容量のメモリを有するメモリ記憶システムにおいて顕著となる。
【０００８】通常使用される他のエラー検出プロセスは、エラー訂正またはエラーチェックおよびエラー訂正コード（以下、ＥＣＣと呼称する）である。一般に、ＥＣＣは、ＣＲＣ（巡回冗長検査）アルゴリズムに基づいている。ＥＣＣコードは、原則としてＣＲＣコードであり、その冗長度は十分に高いので、エラーが破局的でなければ、オリジナルデータを復元することができる。Ｃ

ＲＣアルゴリズムは、データを受信したときに完全なデータシーケンス（データフィールドの終端に付加されたＣＲＣビットを含む）がＣＲＣチェッカによって読み出されるように動作する。完全なデータシーケンスは、ＣＲＣ多項式によって余りなく割切れなければならない。完全なデータシーケンスがＣＲＣ多項式によって割切れない場合、エラーが発生していると判断される。

【０００９】

【発明が解決しようとする課題】パリティに基づく従来からのエラー訂正プロセスと異なり、ＥＣＣコードに基づくシステムは、一般に、複数のビットエラーを検出するために使用できる。例えば、単一ビット訂正を持つＥＣＣメモリシステムは、一般的に、二重ビットエラーを検出し単一ビットエラーを訂正することができる。４または８ビットエラー訂正を持つＥＣＣメモリは、一般的に、４ビットまたは８ビットエラーを検出し訂正することができる。従って、×４または×８構成で編成されたＳＤＲＡＭチップ全体の障害により、システムに障害が発生することはない。ＥＣＣシステムは容易に複数エラービットの検出を行うが、従来からのＥＣＣシステムの問題は、一般に、それらが訂正不可能なエラーを報告する場合にシステムを停止させてしまうということである。このため、最初の停止をしないで、ＥＣＣメモリシステム中の障害部分を置き換えて、システムの障害耐力（immunity：イミュニティ）を回復することはできない。

【００１０】ＥＣＣおよびパリティプロセスは、一般に、週７日、２４時間の動作を要求される、半導体メモリを用いるコンピュータシステムにおいて使用されている。コンピュータシステムが半導体メモリに要求されている速度を必要としない場合か、または半導体メモリを用いた、対コスト上有効でさらに大容量の記憶装置を利用することができない場合は、ディスクドライブメモリが使用される。ディスクドライブメモリによってサポートされるコンピュータシステムは、一般に、ハードウェアによって複数のディスクドライブを互いにリンクさせて、廉価ディスク冗長アレイ（redundant array of inexpensive disks）すなわちＲＡＩＤとして周知の、ドライブアレイを形成する。アレイ中のドライブは、互いに協調しており、データはそれらの間に特別な方法で割り振られる。ディスクドライブおよびディスクドライブを読み出す機械的インタフェースは、半導体メモリより信頼性が低いので、ディスクドライブシステムにおける永続的または一時的なデータ障害のデータ回復プロセスは、一般に、より多くの冗長性とより複雑なデータ回復手順を有している。

【００１１】従来のＲＡＩＤシステムにおいては、データはドライブ間でビットまたはバイトレベルで分割される。例えば、４ドライブシステムにおいて、各バイトの２ビットは第１のハードディスクからもたらされ、次の２ビットは第２のハードディスクからもたらされる、というようにである。従って、４つのドライブは、ドライブを直列にしたときより４倍の速度で単一バイトのデータストリームを出力する。すなわち、１つのドライブが２ビットを転送するために必要な時間と同じ時間しかかからないで、１バイトで構成された情報のすべてが転送される。いくつかのドライブ間でデータを分割するこの技術は、データストライピングと呼ばれ、ドライブ毎の実際のブロックサイズは１０００バイト程度の大きさにすることができる。ＲＡＩＤメモリにより、各ディスクを独立して動作させる場合よりも信頼性が向上し、エラーに対してより大きな抵抗力を持たせることができる。信頼性および障害耐性（fault-tolerance）の向上は、ミラリングおよびパリティの実施を含む種々の冗長方法によって達成される。

【００１２】複雑なデータ回復方法論を必要とせず、複数チャネル上の複数エラーに対する自動的なエラー検出および訂正機能を備えると共に、障害メモリのホットスワッピングもサポートする、メモリシステム構成が必要とされている。

【００１３】

【課題を解決するための手段】本発明は、メモリ障害の後に自己回復（self heal）し、システムのシャットダウンまたはオペレータの介入無しで障害耐力を正常レベルに回復できる、サーバ用の自己回復メモリシステムを提供する。このメモリシステムは、「自己リカバリ」、つまり、アルファ粒子またはＳＤＲＡＭのビットセルウォール障害によってより頻発する単一ビット障害と同様に、ＤＩＭＭにおけるＳＤＲＡＭ障害、ＤＩＭＭコネクタ障害の回復が可能である。

【００１４】本発明の自己回復メモリ構成は、ＳＤＲＡＭ（ＳＤＲまたはＤＤＲ）のような並列メモリ技術またはＲＡＭＢＵＳのような直列メモリ技術と共に使用することができる。好ましい実施の形態では、メモリ構成は、メモリコントローラに電気的に結合されたＣＰＵバスと、スイッチング手段に接続されたメモリコントローラと、複数のメモリモジュールに電気的に接続されたスイッチング手段と、を備えている。ここで、複数のメモリモジュールは、複数のデータメモリモジュールと、複数のＥＣＣメモリモジュールと、ホットスペアメモリモジュールと、を含み、これら複数のデータメモリモジュールとＥＣＣメモリモジュールとホットスペアメモリモジュールとは、並列に接続されている。１または４ビットの訂正可能なハードエラーに対して、エラーが発生したデータメモリモジュールに格納されたデータは、ホットスペアメモリモジュールにマップされる。複数のデータ、ＥＣＣおよびスペアメモリモジュールの並列構成により、ＣＰＵのローディングの低減が容易になり、また必要な場合には、障害が発生したメモリコンポーネントのホットスワッピングが容易となる。

【００１５】好ましい実施の形態は、ホットスペアメモリモジュールを含む。ＥＣＣ　ＤＩＭＭおよびホットスペアを使用するＳＤＲＡＭメモリシステムは、ＳＤＲＡＭの障害またはＤＩＭＭコネクタの障害を原因とする、障害が発生したＤＩＭＭからのデータの再構築を提供する。そして、障害の発生したＤＩＭＭを、オンラインで、あるいはそのシステムがホットスワップをサポートしていない場合は年一度のダウンタイムのうちの数分間に、置換することができる。その後、置換されたＤＩＭＭメモリモジュールは、メモリシステムのホットスペアとなる。

【００１６】本発明の特性および利点は、明細書の残りの部分および添付図面とに関連して更に理解されよう。

【００１７】

【発明の実施の形態】図３は、本発明による自己回復並列メモリ構成３００の部分ブロック図を示す。好ましい実施の形態では、メモリ構成は、メモリコントローラ３１４に電気的に結合されたＣＰＵバス３１０と、スイッチング手段３２０に電気的に結合されたメモリコントローラ３１４と、を備えており、スイッチング手段３２０は、複数のメモリモジュール３２４に電気的に接続されている。複数のメモリモジュール３２４は、好ましくは、複数のデータメモリモジュール３３０と、複数のＥＣＣメモリモジュール３３４と、複数のスペアメモリモジュール３３８と、を有している。好ましい実施の形態では、データメモリモジュール３３０とＥＣＣメモリモジュール３３４とホットスペアメモリモジュール３３８とは、並列に接続されている。

【００１８】図３に示す実施の形態では、データメモリ構成は、１４４ビットのデータメモリチャネル３４０と、３６のメモリモジュールと、を有する自己回復ＤＤＲシステムであり、好ましくは、データメモリモジュール３３０用に３２のＤＤＲ　ＤＩＭＭ、ＥＣＣメモリモジュール３３４用に８つのＤＩＭＭ、および４つのＤＤＲホットスペアＤＩＭＭ３３８を備えている。１４４ビット幅チャネル３４０には、１２８のデータビットと１６のＥＣＣビットとが含まれる。１４４ビットデータバス３４０は、４つの別々の７２ビットバスセグメントに分割される。各バスセグメントは、データに使用される８つのデータメモリモジュール３３０と、ＥＣＣに使用される２つのＥＣＣメモリモジュール３３４と、１つのホットスペアメモリモジュール３３８と、を有しているのが好ましい。ＥＣＣメモリモジュール３３４は、各チャネル３４６ａ～３４６ｄにおいて１６のデータモジュールと４つのＥＣＣモジュールとに渡ってデータストライピングが可能であるように設けられている。

【００１９】データバス、すなわちデータメモリチャネル３４０は、少なくとも１つのスイッチング手段３２０、好ましくは１：２ＦＥＴに電気的に結合されており、メモリコントローラ３１４を介してＣＰＵバス３１０に電気的に結合されている。１：２ＦＥＴスイッチであるスイッチング手段３２０は、１４４ビットバスを２つの１４４ビットバスセグメントに分割する。７２ビット幅ＤＲＡＭの場合、２つの１４４ビットバスセグメントの各々が、２つの別々の７２ビットバスチャネルに分割される。７２ビットメモリチャネルの各々は、複数の７２ビットメモリモジュール３３０，３３４，３３８に並列に接続されている。

【００２０】図３に示す実施の形態において、各メモリチャネルは、２ＥＣＣ　ＤＩＭＭの幅および３６ＤＩＭＭの深さであり、各々１８のＤＩＭＭを持つ２つのバスセグメント（１：２ＦＥＴスイッチによって分離されている）を使用している。各メモリチャネルは、８つのデータＤＩＭＭと、２つのＥＣＣ　ＤＩＭＭと、１つのホットスペアＤＩＭＭと、を有している。従って、メモリコントローラチャネルデータＩ／Ｏピンにおける最大ローディングは、９ＤＩＭＭである。これにより、４つのバスセグメントの各々において、１ＧＢのＤＩＭＭを用いれば３２ＧＢ、または２ＧＢのＤＩＭＭを用いれば６４ＧＢの最大容量およびホットスペアＤＩＭＭが可能となる。

【００２１】図３に示すメモリ構成は、ストライピングを使用する場合、容易に拡大縮小することができない。図３に示すＤＩＭＭの構成および数は、ハイエンドシステムでは一般的であるが、データバスに接続されたＤＩＭＭモジュールの数は、データ記憶用のメモリモジュール１つとホットスペアメモリモジュール１つとであってもよい。データ記憶用メモリモジュールが１つだけ使用される場合は、ＤＩＭＭは、エラー訂正専用の少なくとも１つのメモリ装置またはメモリ装置のビットを有していなければならない。

【００２２】図４は、好ましい実施の形態の自己回復並列メモリ構成の部分ブロック図を示す。図４に示す好ましい実施の形態は、図３に示す実施の形態を縮小したバージョンである。ここに示すメモリ構成は、データまたはＥＣＣのいずれかを格納するために使用できる複数のメモリモジュール３３０と、ホットスペアメモリモジュール３３８ａ，３３８ｂと、を有している。図３に示す実施の形態とは対照的に、専用のＥＣＣ装置は無く、代りに、メモリ装置におけるビットがデータまたはＥＣＣに対して専用化されている。図４に示す実施の形態では、好ましくは、各７２ビットバス３６０ａまたは３６０ｂに提供される単一のホットスペアメモリモジュールがある。メモリモジュール３３０の１つに障害が発生した場合、データはその対応するホットスペアモジュールにマップされる。図４に示す実施の形態では、データストライピングは使用されず、従って、システムは拡大縮小可能である。図４に示す実施の形態では、最低２つのデータメモリモジュールと２つのスペアメモリモジュールとが必要である。

【００２３】図４に示す実施の形態では、データメモリチャネル３４０は、メモリコントローラ３１４を、少なくとも１つのスイッチング手段３２０、好ましくはＦＥＴスイッチに電気的に結合する。ＦＥＴスイッチ３２０は、メモリバスからの信号を分割して、メモリコントローラにおけるローディングを低減する。好ましい実施の形態では、スイッチング手段は、１：２ＦＥＴスイッチである。しかしながら、代替的な実施の形態では、スイッチング手段は、クロスバータイプスイッチとすることができる。しかしながら、システム要件により、特にメモリ構成におけるメモリモジュールの数により、１：２または１：４ＦＥＴを使用することができる。ＦＥＴのサイズの比率が増加するということは、バスセグメントの数が増加することを意味する。

【００２４】図４に示す好ましい実施の形態では、各７２ビットバス３６０ａまたは３６０ｂに対し、ＦＥＴスイッチが使用されている。代替的な実施の形態では、メモリコントローラとＣＰＵバスとの間にスイッチング手段が接続されていない。一般に、わずかな数のメモリモジュールしか必要でない小さいメモリシステム（例えば、１ＧＢレンジ）では、メモリ構成の一部としてスイッチング手段は含まれない。

【００２５】好ましい実施の形態では、メモリモジュール３２４は、ＪＥＤＥＣ準拠を満たすために、７２ビット幅のＤＲＡＭ、好ましくはＤＩＭＭである。しかしながら、システム要件を満たす限りは、いかなるタイプの読出し書込み可能メモリコンポーネントを使用してもよい。メモリモジュール３２４は、スペアであっても稼動中であってもよい。従って、バンク選択、ＲＡＳ等の制御ラインは、可変であり、好ましくは各スロットに対してハードワイヤード（hardwired）されていないことが必要である。

【００２６】好ましい実施の形態では、メモリ構成の各メモリチャネルは、少なくとも１つのデータ訂正用メモリモジュールと、少なくとも１つのホットスペアとして使用されるメモリモジュールと、を有している。少なくとも１つのデータ訂正用メモリモジュールおよび少なくとも１つのホットスペアとして使用されるメモリモジュールは、データバスに並列に電気的に結合されている。好ましくは、メモリモジュールはＳＤＲＡＭであり、システムは、ＳＤＲＡＭ障害全体について訂正を行うことができると共に、ホットスペアにおける再構築プロセス中に発生する可能性のある、追加的なソフトエラーについても訂正を行うことができる。また、制御またはアドレスが冗長な接点を有する場合は、ＤＩＭＭソケット障害についても訂正を行うことができる。

【００２７】メモリ構成は、各メモリチャネルに対し、２つのＥＣＣメモリモジュール３３４を有している。しかしながら、ＥＣＣモジュールの好ましい数は、使用されるＣＲＣ多項式のタイプに限定されず、データエラーの数およびデータエラーが訂正されるチャネルの数を含む、複数の要素によって決まる。メモリコントローラ３１４は、メモリエラーに対しデータ訂正を提供するＥＣＣロジック３４２を含む。ＥＣＣロジック３４２の実現には従来からの技術を用い、それは、使用されるデータ訂正要件およびアルゴリズムによって変更可能である。

【００２８】使用されるＣＲＣ多項式とＥＣＣメモリモジュールの数とは、メモリ構成のデータ訂正要件によって変化する。好ましいＥＣＣの実現では、ＳＤＲＡＭ４つの障害によるエラーを訂正するために１６のＥＣＣビットが必要であり、同時に再構築プロセス中に１つのランダムエラーが許容される。これにより、ビットエラー訂正の数は合計５つとなる。

【００２９】図３に示す実施の形態において、各メモリチャネル３４６ａ～３４６ｄは、ホットスペアメモリモジュール３３８を有している。メモリ部分にホットスペアが存在することを宣言するために、構成ユーティリティプログラムが使用される。ＤＲＡＭ障害全体がＤＩＭＭ（デュアルインラインモジュール）で検出された場合、それは訂正されるが、その後、メモリコントローラ３１４がホットスペアＤＩＭＭ上にデータを再構築する。データビットに対する冗長ビットの割合は、１：８であるが、システムにおける冗長ＤＩＭＭ対ＤＩＭＭの総数の割合により低減される。ホットスペアＤＩＭＭは、少なくとも最大のＤＩＭＭと同程度に大きくなければならない。一般に、ホットスペアメモリモジュールは、他のデータメモリモジュールと同じメモリサイズおよびＤＲＡＭ編成を有している。

【００３０】システム動作中、特定のメモリ位置において十分なエラーが発生した場合、データは、ホットスペアモジュールのメモリ位置に移動される。これにより、不良のデータメモリモジュールか、または異なるデータメモリモジュールの不良のデータメモリ位置が使用不可能となった場合、不良のデータメモリモジュールからのデータは、ホットスペアメモリモジュールに転送される。データメモリモジュール全体またはデータメモリモジュールの大部分が不良となった場合、そのデータメモリモジュールは、最終的にスワップアウトされ、データの完全性が向上しているホットスペアメモリモジュールと交換される。

【００３１】図５から図７は、本発明の自己回復メモリで起こるステップを表す、高レベルブロック図を示す。訂正可能なハードエラーを有するコンピュータシステムに対し障害耐力のレベルを復元するために実行されるステップは、メモリモジュールにおける訂正可能エラーを検出することと、訂正可能エラーが検出されたメモリモジュールからのデータをスペアメモリモジュールに移動して再構築することとを含む。

【００３２】図５は、メモリエラーが第１のメモリモジュールにおいて発生している、図３および図４に示す構

成を表す高レベルブロック図を示す。図５を参照すると、第１のメモリモジュール３３０ａにおいて陰を付けた領域で表された４ビットエラーが示されている。４ビットエラーは、メモリモジュールを読み出す際に検出され複数アドレスで発生する、訂正可能なハードエラーまたは永続的エラーである。また、図５に示すブロック図には、スペアモジュール３３８ａが示されている。

【００３３】図６は、図３および図４に示す構成に対しデータをスペアメモリモジュールに移動して再構築するステップを表す高レベルブロック図を示す。図６を参照すると、矢印３８０は、データメモリモジュール３３０ａからスペアメモリモジュール３３８ａへのデータの移動を表している。移動および再構築プロセス中、データは第１のデータメモリモジュール３３０ａからスペアメモリモジュール３３８ａに転送される。移動および再構築プロセス中は、メモリコントローラによる第１のデータメモリモジュール３３０ａへの新規な書込みはすべて、データメモリモジュール３３０ａとスペアメモリモジュール３３８ａとの両方に対して行われる。矢印３８２は、この二重書込みプロセスを表す。移動および再構築プロセスが完了するまで、読出し動作は第１のデータメモリモジュール３３０ａからのみである。示されている移動および再構築プロセス中は、アルファ粒子により起こるような、一時的な１ビットエラーを訂正することができる。

【００３４】図７は、第１のメモリモジュールからスペアメモリモジュールへのデータの移動および再構築のステップ後の、図３および図４に示す構成を表す高レベルブロック図を示す。図７を参照すると、以前はホットスペアモジュールであったモジュール５は、データモジュールになっている。古い第１のモジュールは、今では第１のモジュールからのオリジナルデータに新たな書込みを足したデータを有している。システムユーザにとって有用である場合には、第１のメモリモジュールは、置換されるべきであり、古い第１のメモリモジュール３３０ａはスペアメモリモジュールとなる。

【００３５】ホットスペアメモリモジュールにデータを移動するために、メモリコントローラにより、不良のメモリ位置とホットスペアのメモリマッピングが行われる。メモリコントローラのデータメモリモジュールへのマッピングを変更することにより（すなわち、不良のＤＲＡＭを有するＤＩＭＭを検出した後の、アドレスまたは制御のルーティングにより）、ホットスペアデータメモリモジュールまたはＤＩＭＭは、書込みおよびメモリ更新のためのホットスペアメモリモジュールと対となるが、読出しは不良のデータメモリモジュールのみから行われるように、設定される。

【００３６】次に、ＣＰＵ、メモリコントローラまたは管理コントローラは、不良のデータメモリモジュールの内容をホットスペアデータモジュールにコピーする。このプロセス中、不良のデータメモリモジュールへのシステム書込みもすべて、ホットスペアメモリモジュールに書き込まれる。従って、コピーが行われると、不良のデータメモリモジュールのイメージに再構築プロセス中に発生したあらゆる書込みを足したものを、ホットスペアは有している。なお、ＤＩＭＭからＤＩＭＭへの転送を行うときは初期化をしないのに対して、すべてのロケーションをコピーする場合は、訂正されたデータを有するホットスペアデータメモリモジュールを初期化することに留意してほしい。

【００３７】エラー訂正方式は、同時に発生する複数のエラーを訂正することができなければならない。このため、例えば、ＤＩＭＭ再構築プロセス中に単一のビットソフトエラーが発生した場合、不良のＤＩＭＭに対し、新たなホットスペアとしてマップアウトする（そしてスワップアウトされるまで不良としてマークする）ことができなければならない。そして、新たなホットスペアＤＩＭＭがメモリ構成内にマップされることにより、不良のＤＩＭＭが完全に置換される。スワップされるべきまたはエラーのロギングのためのＤＩＭＭを識別する際、実際の物理的なＤＩＭＭ位置を計算するためにＤＩＭＭマッパ機能が使用される。

【００３８】例えば、図３において、システムがまだ実行中である時に、不良のデータメモリモジュールをホットスワップアウトするかまたは取出すようにしてもよい。データ転送が行われた後、新たなデータメモリモジュールにより不良のＤＩＭＭを置換することができるように、電源および不良ＤＩＭＭを接続するバス（または取除くべきＤＩＭＭ）がオフにされる。ホットスワッピングが好ましいが、サーバによっては、パッケージングは非常にアクセス密度が高いため、ホットスワッピングが提供されないがまたは実行不可能である場合がある。これらの場合、次に予定されたダウンタイム中に、不良のデータメモリモジュールをメモリ構成からコールドスワップすることができる。

【００３９】好ましい実施の形態では、本発明の実現のために別々のスペアメモリモジュールが必要とされる。スペアメモリモジュールは、メモリ構成においてサポートされなければならない。一般的に、このスペアメモリサポートは、ホットスペアメモリモジュールによって受信されるコマンドおよびアドレスバスが、データモジュールおよびＥＣＣメモリモジュールに送信される信号とは異なっていることが必要である。これは、一般に、データモジュールまたはＥＣＣメモリモジュールにおいてエラーが発生しない限り、ホットスペアメモリモジュールは選択されないためである。訂正可能なハードエラーが発生する場合にのみ、ホットスペアメモリモジュールは起動されることになり、訂正不可能なエラーを有するメモリモジュールは休止状態になる。さらに、このスペアメモリモジュールサポートには、好ましくはメモリコ

ントローラにおいて実現されるか、または代替的ににメモリコントローラ外部の論理回路において実現される、追加ロジックが含まれる。

【００４０】再構築プロセスに対して最も柔軟な方式は、メモリコントローラが、障害の発生したＤＩＭＭを読出し、同時にデータのバーストブロックを（メモリコントローラを経由して）ホットスペアＤＩＭＭに転送する、というものである。障害の発生したＤＩＭＭとホットスペアＤＩＭＭとは、同じメモリサイズおよび編成であるのが好ましいが、障害の発生したＤＩＭＭとホットスペアＤＩＭＭとは、メモリサイズが異なってもよく、異なるＳＤＲＡＭ編成を使用してもよい。再構築プロセス中、システム読出しは障害が発生したＤＩＭＭから行われ、システム書込みは、別々のバーストサイクルで、障害が発生したＤＩＭＭとホットスペアＤＩＭＭとの両方に対して行われる。

【００４１】本発明の１つの実施の形態においては、メモリモジュールに対してデータを書込むために、ジャグラ（juggler）アルゴリズムが使用される。「ジャグラ」アルゴリズムは、同じ棍棒のセットを交互に空中で巧みに操る２人の曲芸師（juggler）に例えてそう名づけられている。ＤＩＭＭデータは、正常なシステムメモリのトランザクションが継続している間に、「オンラインで」不良ＤＩＭＭからホットスペアＤＩＭＭに徐々に移動されるかまたは交互に行われる（juggle）。好ましい実施の形態について、図４に示すメモリ書込みパターンの各ブロックは、８ビットのデータを表しているが、ビットの数は実際のメモリ構成によって変えることができる。例えば、本発明の図３に示す構成における８ビットＤＩＭＭを１６ビットＤＩＭＭモジュールで置換する場合、図４に示す各ブロックは、１６ビットのデータを表すことになる。

【００４２】ＤＩＭＭにおけるＤＤＲ　ＤＲＡＭチップの破局故障か、またはＤＩＭＭコネクタ自体（データピン）の障害の後、システムは、サーバをシャットダウンする必要なく、自身をオンラインで修復し障害耐力の正常レベルに戻る。ジャグラアルゴリズムを用いることで、本発明のメモリ構成は、以下のサーバアプリケーションに適する。すなわち、１）オペレータの居ない遠隔サーバロケーション、２）オンラインでホットスワッピングを行うにはアクセス不可能な程、非常に密度高くパッケージされたメモリシステムを有するサーバ、３）何百または何千のサーバを含むことができる大規模集中型「サーバファーム」において使用されるよう設計され、Ｉ／Ｏ、ＣＰＵおよびメモリサブシステムのホットスワッピングに対し最適化されたサーバ、４）障害のあるサーバを修復のためにシャットダウンすることができ、それゆえ個々のサブシステムにホットスワッピングを提供するという複雑さを要しない、冗長なサーバのペア、である。

【００４３】本発明のメモリ構成において、ＪＥＤＥＣ　ＤＤＲ　ＤＩＭＭ仕様に準拠するように、×４および×８ＳＤＲＡＭに対し１３３ＭＨｚの周波数でエラー訂正が動作する必要がある。ＤＤＲを１００ＭＨｚで使用するデュアルチャネルのバースト帯域幅は、３．２ＧＢ／秒であり、１３３ＭＨｚの場合は４．２ＧＢ／秒である。従って、サーバまたはワークステーションのメモリシステムが、３．２ＧＢ／秒のバースト帯域幅で３：４の割合で１００ＭＨｚ４×ＣＰＵバスと共に使用される場合、非キャッシュコヒーレント４×ＡＧＰメモリアクセスに対し、ワークステーションで追加の帯域幅を使用することができる。あるいはまた、ホットスペアＤＩＭＭ再構築プロセス中に、キャッシュライン転送、Ｉ／Ｏバースト転送、またはメモリコントローラのオーバヘッドに要する、全待ち時間を短縮するため、サーバに帯域幅を追加して使用することができる。ＣＰＵバスが１３３ＭＨｚで動作する場合、メモリは１６６ＭＨｚで動作することができる。ＤＤＲバースト毎に転送されるデータ（バースト長が４とする）は、６４バイトのキャッシュラインまたはＣＰＵバスバーストに対応する。

【００４４】好ましい実施の形態では、検出されたエラーは訂正され記録（ログ）される。イベントログはすべてのエラーを記録し、エラーログは異なるエラークラスに対しておよび特定のＤＩＭＭロケーションに対して行われるのが好ましい。なお、ソフトエラーは一時的な障害であり、ハードエラーは、永続的な障害を表す。このため、ニブル境界における４ビットハードエラーの検出により、ホットスペア上のデータの「再構築」が起動される。同様に、ＤＩＭＭコネクタ障害が検出された場合、優良なＤＩＭＭソケットを有するホットスペア上のデータの再構築を起動するためにそれを使用することができ、あるいは、データを単に訂正することができる。

【００４５】スパイラルな書込みパターンを書込むためには、図３の構成に示すデータメモリモジュールはすべて存在しなければならない。メモリコントローラが異なるメモリアドレスを通過するに従い、データがスパイラルな形式または書込みパターンでデータモジュールおよびＥＣＣメモリモジュールに書込まれる。ＥＣＣとラベル付けされているが、スパイラル書込みパターンの場合、ＥＣＣモジュールはデータ用にも使用され、データモジュールは書込みサイクルの場所によりＥＣＣデータの書込みにも使用される。

【００４６】複雑な方法ではあるが、スパイラル書込みパターンより単純な再構築アルゴリズムを有する代替的な自己回復方法は、非ＥＣＣ　ＤＩＭＭを使用し、ＥＣＣ用に付加された２ＤＩＭＭを含む１６バイトブロックのメモリチャネルに渡ってデータをストライプするというものである。それぞれ２つのホットスペアＤＩＭＭが含まれる場合、これには最低２０のＤＩＭＭが必要である。従って、データビットに対する冗長データビットの

割合は、ホットスペアを含んで１：４である。ＲＡＩＤシステムと同様、ＤＩＭＭはすべて同じサイズでなければならない。

【００４７】データがアレイの各ＤＩＭＭにストライプされると、ブロックサイズが１ニブルかまたは１バイト幅である場合、障害が発生したＤＩＭＭを、障害が発生した時に取り除き置換することができ、また、システムをクラッシュさせることなくＤＩＭＭを取除くことができる。その結果、不良ＤＩＭＭのホットスワップ後、新たなＤＩＭＭは有効なデータを含まなくなり、メモリアクセスが行われるときに再構築することができ、結果として生じるニブルエラーに対する訂正が行われる。あるいはまた、全アドレス空間を、ＣＰＵ、管理コントローラまたはメモリコントローラによってシーケンシャルに読出すことができる。これにより、再構築プロセスが行われるまで，データと共に書込まれる正しいＥＣＣビットを有する新たなＤＩＭＭの各ロケーションに対し、Ｒ－Ｍ－Ｗ（Read-Modify-Write）が行われる。

【００４８】図８は、本発明のメモリモジュールに対してデータを書込むかまたはストライプする第１の方法を示す。図９は、本発明に従ってメモリモジュールに対してデータを書込むかまたはストライプする第２の方法を示す。図８に示す書込みパターンは７２ビット幅のメモリモジュールに適用され、図３に示す実施の形態では、メモリチャネル毎に少なくとも９つのＤＩＭＭと１つのスペアメモリモジュールが必要となる。対照的に、図９に示す書込みパターンは、６４ビット幅のメモリモジュールに適用される。それは、一般的に、図８に示す方法と同じ数のＤＩＭＭが必要であるが、データメモリの容量が少ないため、ストライピングパターンが図８のパターンとは異なっている。

【００４９】データストライピングという代替方法は、（１）ホットスワップ再構築アルゴリズムの複雑さを低減する（不良ＤＩＭＭをいつでもスワップすることができるため）こと、および（２）メモリコントローラとＮＯＳが独立していることが魅力であるが、好ましくないかまたはシステム構成の柔軟性を低下させる可能性のある制約がいくつか追加される。第１に、メモリアレイは、同じサイズのＤＩＭＭ内で完全に占有されなければならない。第２に、必要なＤＩＭＭの数は、４または１８に分割されるＤＩＭＭデータ幅に等しい。これは、ベースシステムにおけるアレイサイズが非常に大きいが、メモリアレイサイズを８ビットチップキルを用いることによって９トータルに縮小することができる、ということを意味する。この代替的な自己回復方式の問題は、データストライピングプロセスがメモリ制御およびアドレス指定の複雑さを増大させるのと同様に電力の消失を増大させる可能性がある、ということである。更に、ＯＬＸによる追加にホットスワップ方式が使用される場合、システムをパワーダウンしメモリアレイデータ全体を再構築することなく、アレイに新たなＤＩＭＭを追加することが複雑になる。これら同じ問題は、ＲＡＩＭと呼ぶこの方式に比較して、ディスクドライブを用いるＲＡＩＤシステムにも存在する。このため、この代替方法は、多くのＤＩＭＭソケットを備えた非常に大容量のメモリを有するハイエンドサーバにより適している。

【００５０】図１０は、ＲＡＭＢＵＳメモリ構成と共に使用することができる本発明の第２の代替的な実施の形態の部分ブロック図を示す。図３に示す実施の形態に比べて、図１０に示す代替的な実施の形態は好ましくない。図１０に示すデイジーチェーン直列アーキテクチャの基本的な問題は、システムの実行中には、データメモリモジュールをバスから取除くこともパワーダウンすることもできず、そのため、ホットスワッピングを行うことができない、ということである。更に、ＥＣＣビットは直列である。従って、１つのコネクタピン障害により、デイジーチェーンにおける１６のバースト的な複数エラーによりシステムに障害がもたらされることとなる。

【００５１】図１０を参照すると、メモリコントローラ４１２に電気的に接続されたＣＰＵバス４１０が示されている。メモリコントローラ４１２は、複数のメモリチャネル４１６ａ，４１６ｂ，４１６ｃに電気的に接続されている。メモリチャネル４１６ａは、ホットスペアメモリモジュールを含んでいる。メモリチャネル４１６ｂは、８つのデータメモリモジュールを含む。好ましくは、メモリチャネル４１６ｃは、３つのＥＣＣメモリモジュールを含む。

【００５２】図１０に示すメモリ構成は、２バイト幅の非ＥＣＣメモリチャネルを使用する。従って、ＲＡＭＢＵＳモジュールの１チャネル分を有するＰＣＢを、システムから取り除くことができる。すなわち、８つのメモリチャネル、２つのＥＣＣメモリチャネルおよび１つのホットスペアチャネルである。ＥＣＣビットの数は、２バイトのチャネル障害を訂正するために十分大きくなければならない。データビットに対する冗長ビットの割合は小さい。従って、図１０に示すメモリ構成は、非常にハイエンドなシステムに対してのみ適している。

【００５３】上記説明は、例示のために示されており、限定することを意図しているのではないことが理解される。例えば、データメモリモジュールからのデータのマッピングを行うためにジャグラアルゴリズムおよびデータストライピングアルゴリズムが示されているが、データをマッピングするために、ＲＡＩＤメモリ構成において一般的に使用されるアルゴリズム等の、他のアルゴリズムを使用してもよい。従って、発明の範囲は、上記説明に関して決定されるべきではなく、添付の特許請求の範囲とかかる請求の範囲が権利を与えている同等物の全範囲に関して決定されるべきである。

【００５４】本発明は例として次の実施態様を含む。

【００５５】（１）　メモリコントローラに電気的に結合されたＣＰＵバスと、メモリコントローラに電気的に結合されたデータバスと、少なくとも１つのデータ記憶用メモリモジュールと、少なくとも１つのスペアメモリモジュールと、を具備し、前記少なくとも１つのデータ記憶用メモリモジュールおよび前記少なくとも一つのスペアメモリモジュールは、前記データバスに並列に電気的に結合されており、前記少なくとも１つのデータ記憶用メモリモジュールで発生する訂正可能なエラーに応答して、前記少なくとも１つのデータ記憶用メモリモジュールからのデータが前記スペアメモリモジュールにマップされるメモリ構成。

【００５６】（２）　前記データ記憶用メモリモジュールは、複数のメモリ装置を含み、前記複数のメモリ装置の少なくとも一部は、エラー訂正に使用される、上記（１）に記載のメモリ構成。

【００５７】（３）　前記少なくとも１つのデータ記憶用モジュールからのデータの前記スペアメモリモジュールへの前記マッピングは、ユーザの介入無しに自動的に発生する、上記（１）に記載のメモリ構成。

【００５８】（４）　少なくとも１つのエラー訂正用メモリモジュールを含み、前記少なくとも１つのエラー訂正用メモリモジュールは、前記少なくとも１つのデータ記憶用メモリモジュールおよび前記少なくとも１つのスペアメモリモジュールに並列に電気的に結合されている、上記（１）に記載のメモリ構成。

【００５９】（５）　前記データバスを少なくとも第１のバスセグメントと第２のバスセグメントとに分割する少なくとも１つのスイッチを含む、上記（１）に記載のメモリ構成。

【００６０】（６）　前記データバスを少なくとも第１のバスセグメントと第２のバスセグメントとに分割する前記少なくとも１つのスイッチがＦＥＴスイッチである、上記（１）に記載のメモリ構成。

【００６１】（７）　ジャグラアルゴリズムを使用して、前記少なくとも１つのデータ記憶用メモリモジュールから前記スペアメモリモジュールにデータがマップされる、上記（１）に記載のメモリ構成。

【００６２】（８）　データストライピングアルゴリズムを使用して、データがデータ記憶用メモリモジュールに書込まれる、上記（１）に記載のメモリ構成。

【図面の簡単な説明】

【図１】　ＤＩＭＭモジュールが直列に接続されているＲＡＭＢＵＳメモリ構成の高レベル部分ブロック図。

【図２】　エラー訂正機能を提供するＲＡＭＢＵＳメモリ構成のデータおよびパリティチャネルの高レベル部分ブロック図。

【図３】　本発明による自己回復並列メモリ構成の部分ブロック図。

【図４】　本発明の好ましい実施の形態による代替的な実施の形態の自己回復並列メモリ構成の部分ブロック図。

【図５】　第１のメモリモジュールでメモリエラーが発生している、図３および図４に示す構成を表す高レベルブロック図。

【図６】　図３および図４に示す構成についてスペアメモリモジュールにデータを移動し再構築するステップを表す高レベルブロック図。

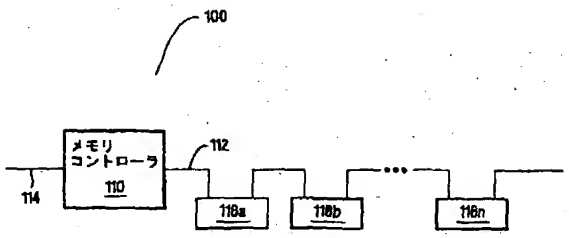
【図７】　第１のメモリモジュールからスペアメモリモジュールへデータを移動し再構築するステップの後の、図３および図４に示す構成を表す高レベルブロック図。

【図８】　本発明による自己回復メモリ構成のメモリモジュールへのデータの書込みまたはストライピングの第１の方法。

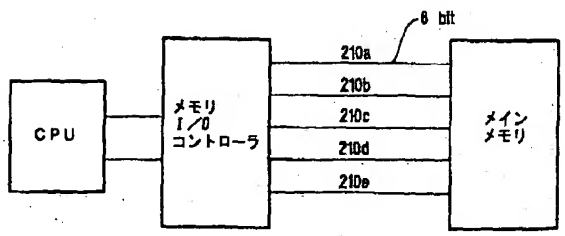
【図９】　本発明による自己回復メモリ構成のメモリモジュールへのデータの書込みまたはストライピングの第２の方法。

【図１０】　ＲＡＭＢＵＳメモリ構成と共に使用することができる本発明の代替的な実施の形態の部分ブロック図。

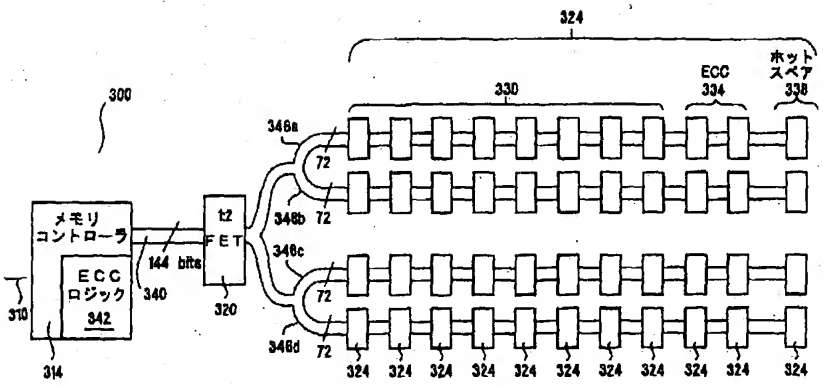
【図１】



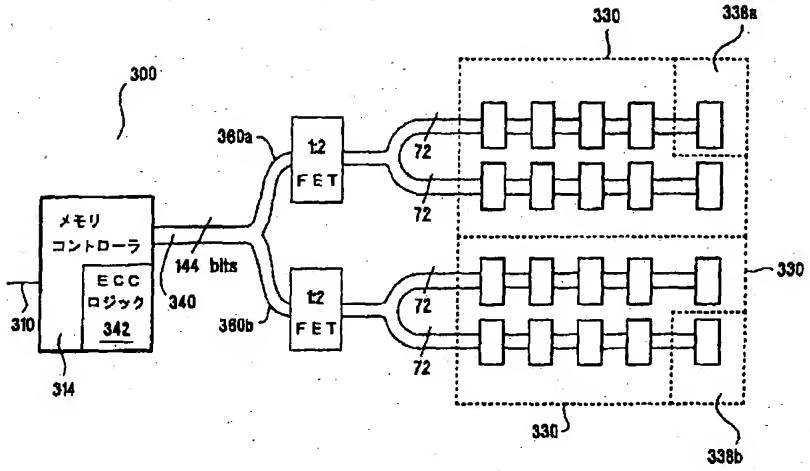
【図２】



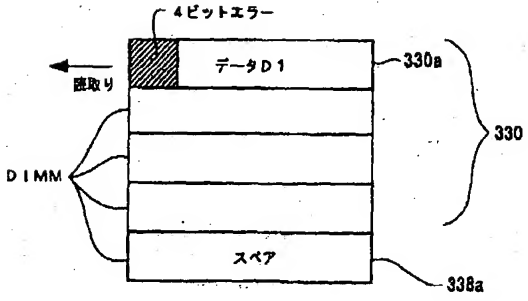
【図３】


324
330
ECC
334
ホット
スペア
338
300
346a
346b
346c
346d
72
メモリ
コントローラ
ECC
ロジック
342
144 bits
340
t:2
FET
320
310
314
324

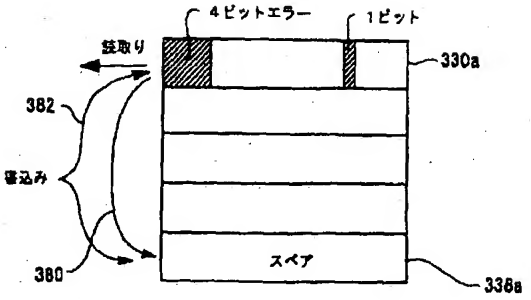
【図４】


330
338a
300
360a
360b
t:2
FET
72
メモリ
コントローラ
ECC
ロジック
342
144 bits
340
310
314
330
330
338b

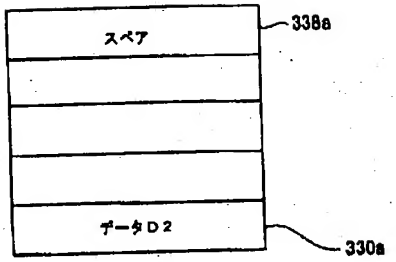
【図５】


4ビットエラー
読取り
データD1
330a
330
DIMM
スペア
338a

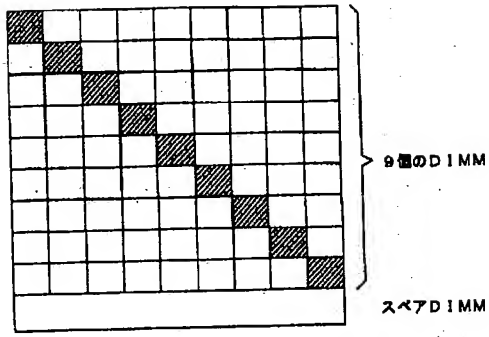
【図６】


4ビットエラー
1ビット
読取り
330a
382
書込み
380
スペア
338a

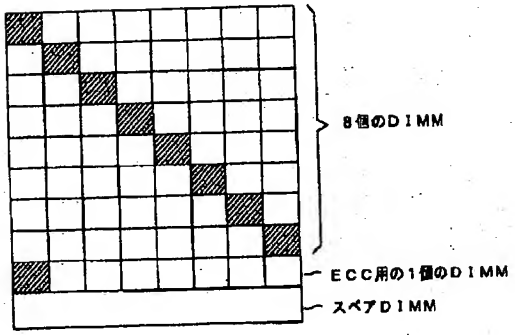
【図７】


スペア
338a
データD2
330a

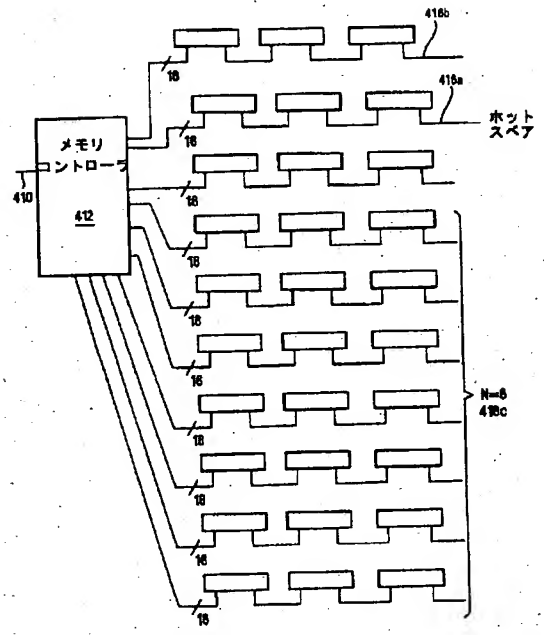
【図８】


9個のDIMM
スペアDIMM

【図９】


8個のDIMM
ECC用の1個のDIMM
スペアDIMM

【図１０】


416b
416a
ホット
スペア
メモリ
コントローラ
410
412
18
N=8
416c